基于改进张量分解模型的个性化推荐算法研究*

陈梅梅　薛康杰
(东华大学旭日工商管理学院　上海　200051)

摘要: 【目的】在基于张量分解的个性化推荐中, 解决因 UGC 标签冗余、热门标签和资源影响用户个性化兴趣所导致的推荐准确性降低问题。【方法】提出一种改进的基于张量分解模型的个性化推荐算法, 引入标签综合共现结合谱聚类的方法, 借鉴 TF-IDF 中 IDF 的思想提出一种基于共现标签和资源的热门惩罚机制, 对基于<用户, 标签簇, 资源>三元关系的初始张量进行重新定义。【结果】基于 Last.fm 数据集的仿真实验结果表明, 从准确率、召回率和 F1 值各项指标上, 本文提出的算法均有良好表现, 综合共现谱聚类的引入使得推荐算法在 F1 值上平均提升 5.91%, 基于 IDF 改进初始张量后的推荐算法在 F1 值上平均提升 1.29%。【局限】未针对其他领域的数据集进行验证, 如微博、Delicious 等。【结论】基于改进的张量分解模型的个性化推荐算法能够显著提高准确性, 有利于社交网络环境下提供更令用户满意的资源。

关键词: 个性化推荐　UGC　标签　标签共现　谱聚类　张量分解
分类号: F224.39　TP391　TP181

1　引　言

随着以 Fackbook、微博为代表的社交网络成为大众维持好友关系及获取信息的主要途径, 基于社交网络的资源推荐已成为时下研究的热门领域, 大量的特征信息可以帮助推荐系统为用户提供更加个性化的推荐。标签作为用户生成内容(User Generated Content, UGC)的一种表现形式, 是基于互联网的社会环境中、由大众用户通过群体智慧形成的一种有效的信息分类、组织和管理方式[1]。用户可以自发地对网络资源进行标注, 通过标签来描述网络资源, 因而, UGC 标签起到了联系用户和资源的纽带作用, 是反映用户兴趣和资源特征的重要数据源。

要融合标签数据, 个性化推荐算法需要充分考虑并完整保留<用户, 标签, 资源>三元关系的特性, 近年来张量分解模型因其对高维数据较好的适应性为基于标签的推荐系统提供了重要的理论支撑[2]。

张量分解模型是矩阵分解模型的高阶推广, 其将三元关系映射到三维矩阵空间, 通过提取主要张量特征值, 得到一个原始张量的压缩近似, 在消除噪声数据的同时能够有效凸显变量之间的隐含关系, 特别适用于解决 UGC 标签存在大量噪声影响推荐准确性的问题, 已经成为基于标签的推荐算法中的主流[2]。常用的张量分解算法主要有 CP 分解和 Tucker 分解[3], 从 1927 年发展至今已较为成熟, 目前研究主要集中在如何针对不同领域的应用进行相应改进。在基于标签的推荐系统方面, Symeonidis 等设计了张量分解在推荐中的通用框架, 并发现高阶奇异值分解(HOSVD)在准确性上要远优于 FolkRank 算法[4]。廖志芳等基于 Tucker 分解和 CP 分解, 提出新用户标签推荐的增量模型[5], 大大降低了推荐算法的时间花费。虽然这些研究成功应用张量分解突出了标签与用户、标签与资源

通讯作者: 陈梅梅, ORCID: 0000-0003-4615-9191, E-mail: cmm@dhu.edu.cn。
*本文系国家社会科学基金项目“中国特色的网络消费调查研究”(项目编号: 10BGL027)的研究成果之一。

之间的关系，但标签的语义模糊及标签冗余问题阻碍基于张量分解的个性化推荐在准确性上进一步提升。为此，有研究从张量模型的构建入手解决这个问题，如 Rendle 等通过对张量内的缺失值进行正负填充，并且优化 AUC 值来获得最优的分解结果[6]；武慧娟等比较标签之间在同一用户同一资源下的优劣性，扩展了三元关系[7]。考虑利用聚类算法对标签数据进行清理，是从问题的核心入手解决标签冗余与语义模糊的有效方案。

另外在推荐问题中热门资源和热门标签也会对推荐结果产生影响，特别是基于社交网络的推荐算法中[8]，由于热门资源往往获得较大的权重导致推荐结果偏向于这些资源而忽略了大量的长尾资源，从而降低了推荐准确性。在二维空间中，一般都通过 TF-IDF 中的 TF 或 IDF 思想来设置惩罚项，以减少热门标签或资源的影响。词频–逆向文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)是用来衡量词能否表达文章特征的方法[9]，其原理是：如果某个词或短语在一篇文档中出现的频率高，并且在其他文档中很少出现，即 TF 高且 IDF 高时，则认为该词或者短语具有很好的类别区分能力。对热门资源的惩罚方面，Fleder 等在商品推荐的研究中，借鉴 TF 思想计算商品间相似度，发现惩罚热门资源有助于提高销售量[10]。王成等在计算用户相似度时，借鉴 IDF 的思想惩罚了热门资源，提高了基于用户的协同过滤算法的准确率和召回率[11]。对热门标签的惩罚方面，Cantador 等发现热门标签无法为区别用户偏好和资源特性提供额外信息，反而降低了推荐准确性[12]，但其忽略了热门资源的影响。项亮将标签作为连接用户和资源的特征，同时惩罚了热门标签和资源，在基于标签的推荐算法准确性上取得了较好的效果[13]。

但在三维空间中当标签被应用于基于张量分解的推荐算法中时，情况却有所不同：Rafailidis 等在对标签聚类后进行初始张量定义时将三元关系拆分成两个二元关系，分别设置惩罚项[14]，这样虽然有助于凸显变化后的三元关系，但相比较凸显三元关系为推荐准确性带来的正面影响，张量本身的高稀疏性导致的负面影响可能更加严重，反而造成推荐准确性的下降。

因此，本文提出一种融合标签综合共现谱聚类和改进的热门惩罚机制的基于张量分解模型的推荐算法。首先，在标签数据预处理中引入基于标签综合共现的谱聚类方法，以在保留三元关系的基础上解决标签语义模糊及冗余问题。其次，针对热门标签和资源影响推荐准确性的问题，在完整保留<用户，标签簇，资源>三元关系基础上，在初始张量重新定义中引入一种改进的惩罚项，从而进一步提升基于张量分解的个性化推荐算法准确性。

2　基于综合共现谱聚类的标签数据预处理

源于大众分类法(Folksonomy)的 UGC 标签存在语义模糊、同义词及多义词问题，会大大降低推荐算法的准确性[15]。在张量分解前，有必要对标签数据进行聚类，以减少标签冗余及语义模糊带来的影响，在消除噪音数据的同时凸显语义关系，有利于提升推荐准确性。

标签聚类即将标签数据分成多个簇，根据标签之间的相似度使得簇内的标签尽可能相似，与其他簇的标签尽可能相异，这样一些不常使用的标签会被一个标签群体所替代，而语义相似的标签也会被归到一个簇中，从而达到凸显用户偏好及资源主题进而提高推荐准确性的目的。

聚类算法通常用于解决稀疏问题，通过选择一个较小的聚类数来达到目的。但针对标签数据的特点，聚类算法能否正确地识别出这些语义模糊及冗余的标签，对于聚类结果合理性影响较大。

Leginus 等对比了几种不同的聚类方法在张量模型下的准确性，发现谱聚类算法要优于其他聚类算法[16]。由于其高效、易于发现不规则聚类的优点，谱聚类算法近年来越来越多地被应用在基于标签聚类的个性化推荐中。对于谱聚类算法来说，标签相似度矩阵是其重要输入之一，如何定义标签之间相似度使其尽可能完整地涵盖标签与用户及资源之间的关系将对最终的标签聚类效果产生重要影响。但 Leginus 等在初始张量的定义中并没有考虑用户标注偏好的差异以及不同资源间标注的差异，因而影响模型的准确度[16]。Symeonidis 在其基础上考虑了用户与标签簇，资源与标签簇的关系，利用向量空间模型计算标签间的余弦相似度，形成相似矩阵再进行谱聚类[17]。

通常标签相似度计算有两种方法：向量空间和标签共现法。向量空间模型将每个标签定义为一个向量，

其中的元素一般代表与用户或者资源与该标签之间的某种关联[18]。但这样的二维向量形式难以表示三维空间的关系。因而, Symeonidis 基于向量空间的标签相似度计算方法无法将用户、标签簇及资源三者关系结合起来看待, 造成了三元关系的分离, 理论上会削弱标签在连接用户与资源语义关系中的重要作用。此外, 在标签向量空间模型中将所有的用户和资源同质化的做法, 随着用户和资源量的快速增长, 向量维度成倍增加, 会造成严重的稀疏性问题从而影响聚类效果。

而基于图论的标签共现法有利于直接表现多元关系。Li 等提出了改进的标签共现结合谱聚类的方法, 将标签相似度分为个体共现相似度和群体共现相似度[19], 个体共现相似度刻画了两个标签间最根本的联系, 而群体相似度增强了标签间的语义关系, 可看作是对个体相似度的补充。其核心观点与李瑞敏等的看法一致, 即如果某资源和某用户之间拥有的共同标签越多, 那么该用户与该资源之间的关联程度越高[20]。通过综合共现相似度将个体和群体共现相似度相结合, 可以更好地表达标签之间的相似关系, 其特点是既不用将三元关系分割成二元组的形式, 又不用将用户资源同质化, 能够在完整地保留用户、标签及资源三者间语义关系的基础上将用户和资源加以区分, 从而帮助聚类算法更好地识别出语义模糊及冗余的标签。

在此基础上本文引入基于综合共现谱聚类方法用于张量分解前对标签数据进行预处理。通过基于综合共现的谱聚类解决标签所固有的语义模糊及冗余问题, 提高分解质量以达到提升推荐准确性的目标。在聚类方法的选择上, 以 K-means 为代表的划分聚类法根据到簇中心的距离进行聚类, 对于某些离簇中心较远的点如果加以修正会在复杂的标签网络中造成巨大的误差。相比之下, 本文标签聚类采用基于图论的谱聚类算法, 即以最小化图权重为目标对图进行切割而形成标签簇, 不存在簇中心, 因而有利于将零散的标签聚合, 同一簇内标签间的相似度都较高且不受距簇中心远近的影响, 较 K-means 算法更利于发现不规则的簇, 从而尽可能降低由于聚类造成的语义损失。

3 初始张量改进

由于张量分解之前进行了标签聚类, 将<用户, 标签, 资源>三元关系转换成<用户, 标签簇, 资源>的形式, 改变了维度定义, 因此需要对初始张量进行适应性改变, 以体现三者之间的相关关系, 同时引入热门惩罚机制以进一步削弱热门标签和热门资源对推荐结果的影响。

在大众分类法中定义一个四元组 F=(U,T,R,Ω)。其中 $\mathrm{U}=\{u_1,u_2,\cdots,u_l\}$ 代表 l 个用户 ID 的集合, $\mathrm{T}=\{t_1,t_2,\cdots,t_m\}$ 代表 m 个标签 ID 的集合, $\mathrm{R}=\{r_1,r_2,\cdots,r_n\}$ 代表 n 个资源 ID 的集合, $\Omega=\{\omega(u_i,t_j,r_k)\,|\,u_i\in\mathrm{U},t_j\in\mathrm{T},r_k\in\mathrm{R}\}$ 代表 u_i 用 t_j 标注 r_k 的可能性集合, 如果有标注记录则 $\omega(u_i,t_j,r_k)=1$, 否则为 0。上述四元组可以转换为张量形式: 定义张量 $B\in R^{I_u\times I_c\times I_r}$, 其中$|I_u|$、$|I_C|$、$|I_r|$分别表示数据集中用户、标签簇和资源的数量, 张量中的元素即 $\omega(u_i,C_j,r_k)$, 通过张量分解算法对初始张量降维, 去除噪音特征值, 获得近似张量 $\hat{B}\in R^{I_u\times I_c\times I_r}$。其中元素为 $\omega'(u_i,C_j,r_k)$ 经过迭代收敛后的值。

通常, 张量内元素表示的是用户、标签和资源三者的关联程度, 当标签聚类成标签簇, 基于<用户, 标签簇, 资源>关系的初始张量内元素可随之改变为:

$$\rho(u_i,C_j,r_k)=\sum_{t_j\in C_j}\omega(u_i,t_j,r_k) \tag{1}$$

公式(1)的含义是将用户 u_i 对资源 r_k 用 C_j 簇内的标签进行标注的次数求和, 作为用户 u_i 对资源 r_k 在 C_j 簇下的权重。

根据公式(1), 若许多用户都使用 C_1 簇中的标签标注了资源 r_1, 即 $\sum_{u_i\in U}\sum_{t_j\in C_1}\omega(u_i,t_j,r_1)$ 较大, 那么系统向用户推荐时, 势必会偏向 C_1 簇下的资源 r_1, 即使用户选择过其他标签簇, 算法也无法较为客观地反映用户个性化的兴趣, 很难发现其他的资源特征。

针对上述问题, 项亮在基于标签的推荐算法中, 提出将 IDF 与对数函数结合来分别惩罚资源和标签的方法[13], 以热门资源为例, 其惩罚项为:

$$\varphi(r_k)=\log(1+n(r_k)) \tag{2}$$

其中, $n(r_k)$ 表示资源 r_k 在不同用户中出现的次数, 如果 $n(r_k)$ 高说明许多用户都有对该资源的标注记录, 以此来说明资源的普遍性, 从而达到识别热门资源的目的。加 1 后取对数可以避免分母为 0 的情况。另外, 由于对数函数相较于线性函数增长较慢, 由此也避免

了分母过大，使得整个公式不易趋近于 0，这也进一步避免了信息缺失，从而较好地解决了热门惩罚的问题。而本文将这一方法引入到三维空间初始张量的定义中。

但是，Gemmell 等在研究中观察到最热门的标签往往含义上较为模糊[15]，导致这些标签可能只在某些含义比较热门，如果单独惩罚标签则可能会对热门标签在非热门含义上造成错误的惩罚。因此本文放弃分别对资源和标签进行惩罚的做法，考虑惩罚共现的热门资源和热门标签，使标签与资源形成对应关系以此确定标签在该资源中的实际含义，避免了由于标签的模糊性导致的错误惩罚。

由此，引入基于共现标签和资源的热门惩罚机制的初始张量定义如下:

$$\omega(u_i,C_j,r_k)=\sum_{t_j\in C_j}\frac{\omega(u_i,t_j,r_k)}{\varphi(t_j,r_k)} \tag{3}$$

其中，$\varphi(t_j,r_k)=\log(1+n_{t_j,r_k})$，$n_{t_j,r_k}$ 表示标签 t_j 和资源 r_k 同时被不同用户使用的次数。可知，如果某个标签总被不同的用户来标注某个资源，则说明该<标签，资源>较为热门，会受到一定的惩罚。而 n_{t_j,r_k} 实际上是标签和资源被使用的交集次数，这样既可以发现并惩罚真正热门的标签及资源，又避免了分开惩罚标签和资源造成的过度惩罚。

4　基于改进张量分解模型的个性化推荐算法

在推荐环节，为了能够尽可能地利用标签簇，发掘用户与资源之间的潜在关系，需要进行张量分解。本文运用 HOSVD-HOOI 算法对初始张量进行分解，选择保留 70%的原始信息[21]，先通过高维奇异值分解算法 HOSVD 去除无用特征值减少张量中的噪声数据以获得一个较好的张量初始解，再运用高维正交迭代 HOOI 算法[22]对初始解进行迭代获得最优近似张量，其中包含三元组之间更为准确的语义关系，可以帮助系统发现用户的潜在兴趣，从而获得更好的推荐。有研究表明相较于其他张量分解算法，这种组合算法能够获得更精确的近似张量[23]。

整个推荐过程从现实使用场景的角度出发，当用户 u_i 点选某个标签 t_j 后，系统会查询 t_j 所属的标签簇 C_j，再找到该用户张量中目标标签簇 C_j 下 $\omega'(u_i,C_j,r_k)$ 最高的 N 个资源，推荐给用户，从而完成推荐。

5　仿真实验

5.1　数据集的选择

本文选用的 Last.fm 数据集①自 2011 年第 5 届推荐系统国际会议发布以来被广泛应用于相关研究，其中包括 2005 年–2011 年间 1 892 名用户对 17 632 位歌手的标注和收听的记录，产生标签 11 946 个，标注行为 186 479 次。为了提高运行效率，对原始数据集进行筛选。首先为避免冷启动问题，选出标注次数大于 70 的用户和歌手；其次，为避免机器人恶意评分影响数据集质量，筛选标注次数小于 3 000 次的用户；最后为避免标签数据过高的稀疏性对聚类效果的影响，筛选出使用次数大于 20 次的标签。最终得到的核心子集包括 444 位用户、275 个标签及 372 位歌手，共 37 749 条有效记录，占总标注次数的 20.24%。从中随机选择 80%的数据作为训练集，剩余的 20%作为测试集。

在训练集中的所有用户都会被随机分配一个自己曾经使用过的标签，通过算法得出每位用户的一个 TopN 列表，通过与测试集中对应用户标注的资源进行比较计算推荐算法相关的性能指标。

5.2　性能评价指标的选择

推荐准确性是评价推荐算法性能的重要指标，目前较为主流的 TopN 推荐结果准确性评价的指标包括：准确率(Precision)、召回率(Recall)及 F1 指标，其中前两者相互影响，因此本文选择使用准确率–召回率曲线定性地反映算法准确性的变化趋势，同时选择作为两者的调和平均数的 F1 指标，以定量地反映算法之间的差距[24]。

仿真实验重复 10 次，对每个指标求其 10 次的均值作为实验结果。

5.3　性能对比

仿真实验模拟用户通过选择一个过去使用过的标签得到一个 TopN 列表的场景，对比从 Top10 到 Top50 每增加 5 个推荐资源时不同算法的性能指标。通过模块度(Modularity Metric)[25]确定最佳聚类个数为 5。

① http://grouplens.org/datasets/hetrec-2011/.

为了检验本文提出的改进的基于张量分解模型的推荐算法(CoSCluIDF)的性能，仿真实验将选择以下三个对比算法:

(1) 基于标签–用户资源矩阵的 K-means 聚类[17]，结合 IDF 初始张量改进的算法方法(KmeansIDF)，以检验本文引入标签共现谱聚类对于基于张量分解模型的推荐算法准确性的提升作用。聚类个数都设为 5，以确保一致。

(2) 综合共现谱聚类结合传统张量分解模型未对初始张量进行任何改进的算法(CoSClu)，以检验本文在初始张量上的改进对于推荐性能的影响。

(3) 仅基于传统张量分解模型不进行任何改进的推荐算法(TD)，以检验本文 CoSCluIDF 算法在张量分解前引入综合共现谱聚类进行数据预处理的做法以及基于改进的热门惩罚机制的初始张量定义对推荐准确性的影响。

5.4 准确性指标对比分析

图 1 显示了准确率–召回率曲线的仿真结果，每条曲线代表一种算法在不同 N 上的准确率和召回率变化。当 N 较小时，准确率较高而召回率较低；随着 N 的变大，准确率下降，召回率上升。曲线的形态越靠右上角，说明推荐效果越好。

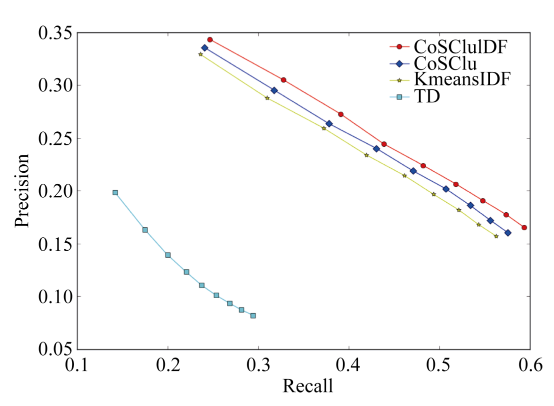


图 1　4 种算法在准确率–召回率曲线上的对比

(1) 本文提出的算法在推荐长度为 10-50 时推荐效果普遍好于另三种算法，与次好的 CoSClu 相比平均准确率相对提升幅度(下同)达 2.69%，平均召回率提升达 2.71%，说明运用综合共现谱聚类并结合 IDF 和基于共现标签的热门标签与资源惩罚机制改进初始张量的做法可以提升推荐准确性。

(2) CoSCluIDF 相较 KmeansIDF 的提升比相较 CoSClu 的提升更为显著，平均准确率提升 5.00%，平均召回率提升达 5.08%。说明合理的标签聚类比惩罚热门标签和资源更有利于推荐准确性的提升。

(3) KmeansIDF 相比 CoSClu 有略微的劣势，平均准确率相差 2.21%，平均召回率相差 2.26%，这也进一步印证了对结果(2)的分析，即综合共现谱聚类对于推荐准确性的提升比惩罚热门标签和资源更为明显。

图 2 为 4 种算法在 F1 指标上的对比。

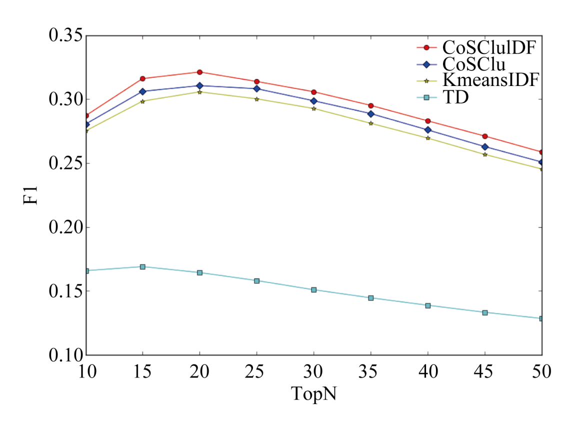


图 2　4 种算法在 F1 指标上的对比

(1) 本文提出的 CoSCluIDF 与 KmeansIDF 相比，在 F1 指标上平均提升达 5.04%，最大提升达 5.91% (N=15)，说明综合共现谱聚类相比传统的 K-means 方法更有利于提升推荐的准确性。

(2) CoSCluIDF 相比 CoSClu，在 F1 指标上平均提升 1.29%，最大提升 1.90%(N=35)，说明引入 IDF 思想定义初始张量并基于共现标签和资源的热门惩罚机制的改进，有效消除了热门标签和资源对降低算法准确性的影响。

通过图 1 与图 2 对比发现：相较 TD 算法，CoSClu、KmeansIDF 和 CoSCluIDF，无论是准确率–召回率曲线还是 F1 指标上性能都有大幅提升。说明稀疏性对于推荐准确性的影响最为明显，聚类算法通过设定一个较小的聚类数能有效解决稀疏问题；但由于维度相同，当聚类数一样时，不同聚类算法对于稀疏性的解决效果也几乎相同，但在更好地识别标签冗余及语义模糊以提升推荐准确性方面，综合共现谱聚类相对更有优势。

另外在实际应用中，尤其是在用户已经标注了一些资源的情况下，很少会只给用户推荐 10 个或者更低的资源，而是将大于某个阈值的所有资源按照权重从大到小排序，全部推荐给用户。因此可以认为较大的

推荐长度更具有现实意义。由图 2 中的仿真结果对比发现: 在 N=20 时, CoSCluIDF 算法在 F1 上能够获得最大值, 具有实际意义。因此最理想的推荐长度建议为 N=20。

为了避免仿真结果受标签筛选的影响, 本文特在其他筛选条件不变情况下对比标签出现次数大于 8 次所得到的核心子集下(简称为 Tag8)的仿真结果。不同数据集下各个算法在三个准确性指标上的均值表现具体数据如表 1 所示。

表 1 推荐平均准确性指标对比

数据集	准确性指标	CoSCluIDF	CoSClu	KmeansIDF	TD
Tag8	Precision	**22.69%**	22.56%	22.38%	11.39%
	Recall	**43.61%**	43.36%	42.94%	21.32%
	F1	**28.21%**	28.05%	27.80%	13.99%
Tag20	Precision	**23.67%**	23.05%	22.54%	12.20%
	Recall	**45.80%**	44.59%	43.59%	23.05%
	F1	**29.48%**	28.71%	28.07%	15.03%

可以看到稀疏性对所有算法的准确性均有影响, 但是, 即便在高稀疏性数据情况下, 本文提出的 CoSCluIDF 相对于其他算法在各项性能指标上都表现最佳, 只是相对于 Tag20 数据集较低稀疏性情况下, 性能优势略有削弱。可见, 高稀疏性标签数据中一些杂乱标签会影响综合共现谱聚类对合理标签簇的发现能力, 因此建议在实际应用中尽可能先对标签数据进行清理, 以减小稀疏性对算法的影响。

6 结 语

为优化基于 UGC 标签的个性化推荐结果的准确性, 本文在张量分解模型中引入标签综合共现结合谱聚类的方法通过保留<用户, 标签, 资源>三元关系的语义完整性来有效识别相似的标签, 缓解标签冗余及语义模糊对推荐准确性的影响。进而, 为了解决三维空间上热门标签和资源对推荐准确性的影响, 在 TF-IDF 中的 IDF 思想上提出一种基于共现的标签和资源的热门惩罚机制并在此基础上重新定义了初始张量, 既保留了三元语义关系又能凸显用户的个性化兴趣。仿真实验表明这种方法能够充分利用标签数据信息, 有效提高推荐算法的性能。

随着在线社交网站的普及, 基于标签和信任关系的个性化推荐将会受到更加广泛的关注。未来工作将集中在进一步充分利用标签簇和社交网络信任关系信息的基于张量分解模型推荐算法性能优化研究。

参考文献:

[1] Moens M F, Li J, Chua T S. Mining User Generated Content[M]. CRC Press, 2014: 7-9.

[2] Marinho L B, Nanopoulos A, Schmidt-Thieme L, et al. Social Tagging Recommender Systems[M]. USA: Springer US, 2011: 615-644.

[3] Hitchcock F L. The Expression of a Tensor or a Polyadic as a Sum of Products[J]. Journal of Mathematics & Physics, 1927, 6(1): 164-189.

[4] Symeonidis P, Nanopoulos A, Manolopoulos Y. Tag Recommendations Based on Tensor Dimensionality Reduction[C]//Proceedings of the 2008 ACM Conference on Recommender Systems, Lausanne, Switzerland. ACM, 2008: 43-50.

[5] 廖志芳, 王超群, 李小庆, 等. 张量分解的标签推荐及新用户标签推荐算法[J]. 小型微型计算机系统, 2013, 34(11): 2472-2476. (Liao Zhifang, Wang Chaoqun, Li Xiaoqing, et al. Tag Recommendation and New User Tag Recommendation Algorithms Based on Tensor Decomposition[J]. Journal of Chinese Computer Systems, 2013, 34(11): 2472-2476.)

[6] Rendle S, BalbyMarinho L, Nanopoulos A, et al. Learning Optimal Ranking with Tensor Factorization for Tag Recommendation[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 727-736.

[7] 武慧娟, 徐宝祥, 王艳艳. 基于张量分解的个性化信息推荐方法优化研究[J]. 情报科学, 2014, 32(6): 134-137. (Wu Huijuan, Xu Baoxiang, Wang Yanyan. Optimization Research of Personalized Tag Recommendation Method Based on Tensor Decomposition[J]. Information Science, 2014, 32(6): 134-137.)

[8] Celma S, Cano P. From Hits to Niches? or How Popular Artists Can Bias Music Recommendation and Discovery[C]// Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition, Las Vegas, Nevada. ACM, 2008: 1-8.

[9] Salton G, Buckley C. Term-weighting Approaches in Automatic Text Retrieval[J]. Information Processing & Management an International Journal, 1988, 24(5): 513-523.

[10] Fleder D, Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales

Diversity[J]. Management Science, 2007, 55(5): 697-712.

[11] 王成, 朱志刚, 张玉侠, 等. 基于用户的协同过滤算法的推荐效率和个性化改进[J]. 小型微型计算机系统, 2016, 37(3): 428-432. (Wang Cheng, Zhu Zhigang, Zhang Yuxia, et al. Improvement in Recommendation Efficiency and Personalized of User-based Collaborative Filtering Algorithm [J]. Journal of Chinese Computer Systems, 2016, 37(3): 428-432.)

[12] Cantador I, Bellogín A, Vallet D. Content-based Recommendation in Social Tagging Systems[C]// Proceedings of the 4th ACM Conference on Recommender Systems, Barcelona, Spain. ACM, 2010: 237-240.

[13] 项亮. 推荐系统实践[M]. 人民邮电出版社, 2012: 107-108. (Xiang Liang. Practice of Recommendation System[M]. Posts & Telecom Press, 2012: 107-108.)

[14] Rafailidis D, Daras P. The TFC Model: Tensor Factorization and Tag Clustering for Item Recommendation in Social Tagging Systems[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2013, 43(3): 673-688.

[15] Gemmell J, Ramezani M, Schimoler T, et al. The Impact of Ambiguity and Redundancy on Tag Recommendation in Folksonomies[C]//Proceedings of the 3rd ACM Conference on Recommender Systems, New York. ACM, 2009: 45-52.

[16] Leginus M, Dolog P, Žemaitis V. Improving Tensor Based Recommenders with Clustering[C]//Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization, Montreal, Canada. Springer-Verlag, 2012: 151-163.

[17] Symeonidis P. ClustHOSVD: Item Recommendation by Combining Semantically Enhanced Tag Clustering with Tensor HOSVD[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, 46(9): 1-12.

[18] Shepitsen A, Gemmell J, Mobasher B, et al. Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering[C]//Proceedings of the 2008 ACM Conference on Recommender Systems, Lausanne, Switzerland. ACM, 2008: 259-266.

[19] Li H, Hu X, Lin Y, et al. A Social Tag Clustering Method Based on Common Co-occurrence Group Similarity[J]. Frontiers of Information Technology & Electronic Engineering, 2016, 17(2): 122-134.

[20] 李瑞敏, 林鸿飞, 闫俊. 基于用户–标签–项目语义挖掘的个性化音乐推荐[J]. 计算机研究与发展, 2014, 51(10): 2270-2276. (Li Ruimin, Lin Hongfei, Yan Jun. Mining Latent Semantic on User-Tag-Item for Personalized Music Recommendation[J]. Journal of Computer Research and Development, 2014, 51(10): 2270-2276.)

[21] Symeonidis P, Nanopoulos A, Manolopoulos Y. A Unified Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(2): 179-192.

[22] Lathauwer L D, Moor B D, Vandewalle J. On the Best Rank-1 and Rank-(R1, R2,…, RN) Approximation of Higher-Order Tensors[J]. Siam Journal on Matrix Analysis & Applications, 2000, 21(4): 1324-1342.

[23] Kolda T G, Bader B W. Tensor Decompositions and Applications[J]. College & Research Libraries, 2005, 66(4): 294-310.

[24] Pazzani M, Billsus D. Learning and Revising User Profiles: The Identification of Interesting Web Sites[J]. Machine Learning, 1997, 27(3): 313-331.

[25] White S, Smyth P. A Spectral Clustering Approach to Finding Communities in Graph[C]//Proceedings of the 2005 SIAM International Conference on Data Mining, Newport Beach, CA, USA. SIAM, 2005: 274-285.

作者贡献声明:

陈梅梅: 提出研究思路和具体研究方案, 论文撰写与修订;
薛康杰: 方法设计与仿真实现, 论文撰写与修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: cmm@dhu.edu.cn。

[1] 薛康杰. train.txt. 训练数据集.

[2] 薛康杰. test.txt. 测试数据集.

收稿日期: 2016-11-10
收修改稿日期: 2017-01-16

Personalized Recommendation Algorithm Based on Modified Tensor Decomposition Model

Chen Meimei　Xue Kangjie
(Glorious Sun School of Business & Management, Donghua University, Shanghai 200051, China)

Abstract: [**Objective**] This paper tries to improve the prediction accuracy of personalized recommendation algorithm based on the tensor decomposition model. [**Methods**] First, we proposed a new tensor model using spectral clustering technique based on combined tag co-occurrence. Second, we established a penalty scheme on popular tag and resource co-occurrence with the help of IDF in TF-IDF. Finally,we re-defined the initial tensor on the triplets of user, tag cluster, and resource. [**Results**] We examined the proposed model with dataset from Last.fm and found its precision, recall and F1 measure outperformed other algorithms. The F1 measures were increased by 5.91% and 1.29% thanks to the two proposed modifictions based on clustering and IDF. [**Limitations**] The proposed algorithm should be further evaluated with datasets from Weibo, Delicious, and other resources. [**Conclusions**] The new algorithm based on advanced tensor decomposition model could significantly improve the accuracy of resources recommendation to satisfy social network system users' information needs.

Keywords: Personalized Recommendation　UGC　Tag　Tag Co-occurrence　Spectral Clustering　Tensor Decomposition

---

VitalSource 和加州州立大学合作改善开放教育资源的利用情况

开放教育资源(Open Educational Resources, OER)在过去 10 年中备受欢迎，其为学生提供了经济实惠还可以轻松定制以满足导师个性需求的课程材料。然而，这些学习材料通常缺乏关键功能，例如可靠的分发、简单的集成和详细的分析。

为弥合开放教育资源和传统学习材料之间的差距，加州州立大学(CSU)集团和 VitalSource 技术公司于近日宣布合作，旨在改善开放教育资源的采用和使用情况。VitalSource 技术公司是 Ingram 内容集团旗下数字教育内容交付方面的全球领导者。

VitalSource 副总裁 Mike Hale 说: "这一合作与我们的使命是一致的，目的是帮助创建和交付价格合理、高质量的课程资料。教授和指导老师正在投入大量的时间和精力来创建开放教育资源。这一合作能促使优质的开放教育资源内容能够具有和 VitalSource Bookshelf 平台同等水平的可发现性、易用性，市场覆盖度，以及平台可靠性。"

加州州立大学管理的 MERLOT、SkillsCommons 和 COOL4Ed 项目中现有许多学术和职业发展开放教育资源，有兴趣采用这些资源的教育者和机构，将能够通过 VitalSource Bookshelf 平台向教师和学生提供这些内容。希望创建或修改开放教育资源内容的教育者和机构也可以继续使用 VitalSource Content Studio 平台和 VitalSource 专有的数字创作工具。该工具为内容创作者提供了直观的操作，能创建基于标准的响应式的、交互式的和可访问的内容。在 VitalSource Content Studio 中创建的内容可以通过 Bookshelf 平台分发给学生。

"这一合作将使得个人和机构能够方便地、可扩展地、可持续地使用开放教育资源，" 加州州立大学副校长 Gerry Hanley 说，"未来，我们将有一个数字化的市场，为教育工作者和学习者提供最便宜的教育内容，并提供方便可靠的分发服务。"

(编译自: http://press.vitalsource.com/oer-adoption-made-easy-through-vitalsource-and-california-state-university)

(本刊讯)